Box-Cox変換と不完全ガンマ関数を用いた成長曲線に関する研究

Box-Cox 変換と不完全ガンマ関数を用いた成長曲線に関する研究

A STUDY ON A GROWTH CURVE USING BOX-COX TRANSFORMATION AND INCOMPLETE GAMMA FUNCTION

安部彰悟
Shogo ABE
指導教員　木村 光宏

法政大学大学院理工学研究科システム理工学専攻修士課程

In the time series data analysis, growth curves have been playing an important role to forecast the future trend of the dataset for example. Among such curves, Gomperz and logistic ones are widely-known formulas in the literature. In this study, based on the recent developments for the modeling, we propose a unified expression of several kinds of growth curves by utilizing the Box-Cox transformation. The obtained formula consists of the incomplete gamma function and provides high flexibility. We show the method of parameters estimation and the actual data analysis are also provided. As a result of the model verification, we found that the fitting performance of our model was improved than the existing models.

***Key Words** : growth curve, Box-Cox transformation, incomplete gamma function*

1. はじめに

(1) 研究背景

古くから，ソフトウェアの累積発見フォールト数やコンピュータウイルスの感染数など, 序盤は少なく途中で急激に増加し，終盤収束していく S 字曲線を描く現象に対して，各データセットに沿うような様々な成長曲線を利用して，将来予測をする研究やデータセットに対してより当てはまりが良い新しい成長曲線を提案する研究が数多くなされている [1]. 数多くの成長曲線が提案される理由として，モデルが適用例として挙げるデータセットに対してはうまく当てはまるのに対し，それ以外のデータセットではうまく当てはまらないということが頻繁に生じることが挙げられる．

(2) 研究目的

本研究では，Box-Cox 変換 [2] と不完全ガンマ関数で与えられる関数に着目し，いくつかの知られた成長曲線を統一的に表現する手法を提案する．また，実測データを用いて 95% 信頼区間の導出やデータの早期推定などの数値実験を行い，得られた結果からモデルの有用性などについて検討する．

2. データセット

本研究で扱うデータ (DS-1, DS-2) を表 1, 表 2 に示す．DS-1 はある年の 9 月末から翌年の 11 月 14 日までの累積バグ発見数のデータで，第 i 週目の時間を x_i とし，1 週目は 0.1 としている．また，第 i 週目の累積発見バグ数を y_i とする [3]．本研究で使用する際には，第 i 週目を $t_i(i = 0, \cdots, 58)$，累積発見バグ数を $y_i(i = 0, \cdots, 58)$ とした．また後述するモデルの都合上,1 週目の累積発見バグ数 248 を 1 とし，すべての累積発見バグ数の値を 247 引いたものを使用した．DS-2 は国内のソフトウェア会社から得た実測データで，テスト工程の進捗率に対するバグの累積発見数が与えられている．分析する際には，進捗率の 5 %を t_1 として，$t_1 = 1, y_1 = 47$ とし，同

表 1 加工した DS-1.

時間 t_i	累積バグ発見数 y_i	時間 t_i	累積バグ発見数 y_i
0	1.	30	4104.
1	15.	31	4154.
2	125.	32	4192.
3	279.	33	4241.
4	495.	34	4301.
5	711.	35	4349.
6	968.	36	4382.
7	1224.	37	4433.
8	1491.	38	4466.
9	1689.	39	4502.
10	1724.	40	4536.
11	1900.	41	4570.
12	2011.	42	4602.
13	2171.	43	4630.
14	2320.	44	4654.
15	2441.	45	4681.
16	2562.	46	4703.
17	2678.	47	4723.
18	2779.	48	4751.
19	2958.	49	4777.
20	3101.	50	4813.
21	3229.	51	4838.
22	3326.	52	4841.
23	3472.	53	4843.
24	3503.	54	4863.
25	3705.	55	4882.
26	3801.	56	4892.
27	3890.	57	4920.
28	4004.	58	4939.
29	4054.		

表 2 加工した DS-2.

時間 t_i	実績値 y_i
0	1
1	48
2	94
3	123
4	160
5	185
6	202
7	206
8	226
9	239
10	263
11	279
12	298
13	306
14	311

様に $t_{14} = 14, y_{14} = 310$ まで置き換えた．また，$t_0 = 0, y_0 = 1$ を追加し，y_1 から y_{14} の値に 1 を加えてモデルを使用できるように調整した．

3. モデル構築の準備

(1) Box-Cox 変換

一般に，Box-Cox 変換は，観測された正の値を取るデータの散布の様子を，正規分布に近づけるために用いられる変換として知られ，あるデータ点 $y(> 0)$ に対してパラメータ λ を導入した Box-Cox 変換値 $y^{[\lambda]}$ は，以下の式で与えられる．

$$y^{[\lambda]} = \begin{cases} \log y & (\lambda = 0) \\ \frac{y^\lambda - 1}{\lambda} & (\lambda \neq 0) \end{cases}. \tag{1}$$

式 (1) において，$\lim_{\lambda \to 0}(y^\lambda - 1)/\lambda = \log y$ となるので，Box-Cox 変換は対数変換の拡張と考えることもできる．

(2) 数値差分法

本研究では，微分係数をデータの値を用いた数値微分で近似する際に差分法を用いる．データが (t_i, y_i) で与えられていたとすると，前進差分の場合の近似式は，

$$\frac{dH(t)}{dt}\Big|_{t \to t_i} = \frac{y_{i+1} - y_i}{t_{i+1} - t_i},$$

となる．同様に，後退差分の近似式は

$$\frac{dH(t)}{dt}\Big|_{t \to t_i} = \frac{y_i - y_{i-1}}{t_i - t_{i-1}},$$

となり，中心差分の近似式は

$$\frac{dH(t)}{dt}\Big|_{t \to t_i} = \frac{1}{2}\left(\frac{y_i - y_{i-1}}{t_i - t_{i-1}} + \frac{y_{i+1} - y_i}{t_{i+1} - t_i}\right),$$

となる．

4. 成長曲線の定式化

$t \geq 0$ として，非負の狭義単調増加関数 $H(t)$ について以下の微分方程式を考える [1],[4].

$$\log\left[\frac{dH(t)}{dt}/t^c/H(t)^d\right] = A - Bt, \tag{2}$$

ここで，A, B, c, d はそれぞれ定数パラメータである．$B > 0$, $c > -1, d \neq 1$ のとき，この式を $H(0) = h_0(> 0)$ として解くと以下の式が得られる．

$$H(t) = \left[\frac{h_0^{d-1}}{1 - (d-1)\frac{e^A}{B^{c+1}}\gamma(c+1, Bt)h_0^{d-1}}\right]^{-\frac{1}{1-d}}. \tag{3}$$

式（3）を変形すると

$$\frac{H(t)^{1-d} - h_0^{1-d}}{1-d} = \frac{e^A}{B^{c+1}}\gamma(d+1, Bt), \tag{4}$$

が得られ，$h_0 = 1$ とすると，式（4）の左辺が $\lambda = 1 - d$ の場合の Box-Cox 変換の式になるので，式 (1) を用いることで次の式が得られる．

$$H(t)^{[1-d]} = \frac{e^A}{B^{c+1}}\gamma(c+1, Bt). \tag{5}$$

ただし，d は有理数であることに注意を要する．ここで，$\gamma(c+1, Bt)$ は第 1 種の不完全ガンマ関数

$$\gamma(c+1, Bt) = \int_0^{Bt} s^c e^{-s} ds,$$

である．$H(t)$ の具体例を示すと，$c = 0, d = 0$ のとき，$H(t)$ はソフトウェアの信頼性評価モデルで用いられる指数形の成長曲線 [5] を表し，

$$H(t) = 1 + \frac{e^A}{B}(1 - e^{-Bt}),$$

となる．$c = 0$，$d = 1$ のとき，式 (1) を経由して求めると $H(t)$ はゴンペルツ曲線 [5] を表し，

$$H(t) = e^{\frac{e^A}{B}(1 - e^{-Bt})},$$

となる．同様にして $c = 1$，$d = 0$ のとき，$H(t)$ は前述の指数形と並んでソフトウェア信頼性評価モデルでよく現れる遅延 S 字形の成長曲線 [5] を表し，以下のように与えられる．

$$\begin{aligned} H(t) &= 1 + \frac{e^A}{B^2}\gamma[2, Bt] \\ &= 1 + \frac{e^A}{B^2}(1 - (1 + Bt)e^{-Bt}), \end{aligned}$$

となる．さらに $c = 0$，$d = 2$ のとき，$H(t)$ はロジスティック曲線 [5] を表し，

$$H(t) = \frac{1}{1 - \frac{e^A}{B}(1 - e^{-Bt})},$$

となる．このように，式 (5) は信頼性評価でよく用いられるモデルの統一式になっている．また，$H(0) = 1$ としたため，データを分析する際には，データの初期値が 1 になるように縦に平行移動させる必要がある．

5. パラメータ推定法

本研究のモデリングと推定では，2 通りのパラメータ推定法が考えられる．

(1)　データに曲線を直接当てはめることによるパラメータ推定

実測データ (t_i, y_i) に対して，$d \neq -1$ として

$$H(t) = \left[1 + (1-d)\frac{e^A}{B^{c+1}}\gamma(c+1, Bt)\right]^{\frac{1}{1-d}}, \qquad (6)$$

を直接当てはめ，データとの誤差が最小となるような A, B, c, d を推定する．この方法でパラメータを推定するとデータに対して $H(t)$ がよく当てはまるが，初期値を正確に与えなければパラメータを推定することができない．また，$\hat{A}, \hat{B}, \hat{c}, \hat{d}$ が従う分布が分らないため，信頼区間などを求めることができない．

(2)　単回帰モデルによるパラメータ推定

式 (2) の関係式を用いて，左辺を差分法を用いてデータから数値微分により与えることで，未知パラメータ c, d を含んだ式（2）の左辺の値が与えられる．ここで，$z(c,d,t_i) = \log\left[\frac{dH(t)}{dt}/t^c/H(t)^d\right]|_{t=t_i}$ とし，

$$z(c,d,t_i) = A - Bt_i, \qquad (7)$$

の単回帰モデルとして A, B, c, d を推定する．このとき，式 (7) の右辺には真の回帰直線との誤差項 $\varepsilon_i \sim N(0, \sigma^2)$ を想定し，

$$z(c,d,t_i) = A - Bt_i + \varepsilon_i, \qquad (8)$$

と考えていることになる．式 (8) には未知パラメータが 4 つ含まれているが，最小二乗法の下で，$\hat{A}, \hat{B}$ が c, d を含んだ以下の式

$$\hat{A} = \frac{1}{n}\sum_{i=0}^{n} z(c,d,t_i) - \frac{\hat{B}}{n}\sum_{i=0}^{n} t_i$$

$$\hat{B} = \frac{\sum_{i=0}^{n} t_i z(c,d,t_i) - \frac{1}{n}\sum_{i=0}^{n} z(c,d,t_i)\sum_{i=0}^{n} t_i}{\sum_{i=0}^{n} t_i^2 - \frac{1}{n}\left(\sum_{i=0}^{n} t_i\right)^2}, \qquad (9)$$

で求めることができるため，モデルの自由度は 2 となっていることに注意が必要である．式 (8) の誤差二乗和

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \{z(c,d,t_i) - (\hat{A} - \hat{B}t_i)\}^2, \qquad (10)$$

を最小とする $\hat{c}, \hat{d}$ を求めて，再度式 (9) を経て $\hat{A}, \hat{B}$ を求めることができる．この方法でパラメータを推定すると，式 (8) の最小二乗法の帰結として，$\hat{A}, \hat{B}$ が正規分布に従うため，新しい時点 $t_f (t_f > t_n)$ における $z(t_f)$ の信頼区間を構成することができる．ただ，これらの推定値を $H(t)$ に代入しグラフを描くと，図 1 のようにデータの散布と $H(t)$ に乖離が生じる．

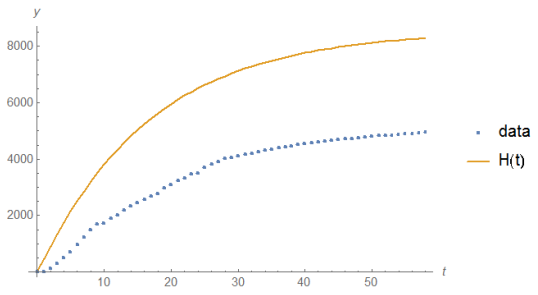

図 1　回帰のパラメータを用いた DS-1 の $H(t)$.

そのため，本研究では，式 (9)，式 (10) で求めた推定値を初期値として用いて，実測データと式 (6) との誤差二乗和

$$\sum_{i=1}^{n}\left[y_i - \left\{1 + (1-\hat{d})\frac{e^{\hat{A}}}{\hat{B}^{\hat{c}+1}}\gamma(\hat{c}+1, \hat{B}t)\right\}^{\frac{1}{1-\hat{d}}}\right]^2, \qquad (11)$$

を最小にするパラメータ A, B, c, d を求めるという 2 段階のステップを踏んで，$H(t)$ を求めていく．

6.　回帰直線の 95% 信頼区間と予測区間

$y_i = \alpha + \beta x_i + \varepsilon_i$ として，Y_i と残差 ε_i が正規分布に従っているとき，a を α の不偏推定量，b を β の不偏推定量，$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x$，$S_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$，$V_e = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ とすると $\alpha + \beta x$ の 95% 信頼区間は

$$a + bx - t_{0.025}(n-2)\sqrt{\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{S_x}\right)V_e} \leq \alpha + \beta x$$

$$\leq a + bx + t_{0.025}(n-2)\sqrt{\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{S_x}\right)V_e}, \qquad (12)$$

となる．同様に $\alpha + \beta x$ の 95% 予測区間は

$$a + bx - t_{0.025}(n-2)\sqrt{\left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_x}\right)V_e} \leq \alpha + \beta x$$

$$\leq a + bx + t_{0.025}(n-2)\sqrt{\left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_x}\right)V_e}, \qquad (13)$$

となる [6]．本研究では，いくつかの仮定の下でこのことを用いて，データの将来予測などを行う．

7.　数値例

(1)　$H(t)$ のパラメータ推定

実際にデータを用いて，$H(t)$ のパラメータを求める．その際に，本研究では前進差分，後退差分，中心差分の 3 つの方法を用いて近似を行い，結果を比較していく．

a)　前進差分の場合

$z(c,d,t_i)$ を前進差分を用いて，以下のように変形する．ただし，$i = n$ の場合は前進差分を行えないので，代わりに後退差分を用いる．

$$z(c,d,t_i) = \begin{cases} \log[\frac{y_{i+1}-y_i}{t_{i+1}-t_i}/t_i^c/y_i^d] & (1 \leq i < n) \\ \log[\frac{y_n-y_{n-1}}{t_n-t_{n-1}}/t_n^c/y_n^d] & (i = n) \end{cases}. \qquad (14)$$

式 (14) を用いて，式 (9)～(11) を計算し，パラメータを推定した結果を表 3, 表 4 に，推定したパラメータを $H(t)$ に代入してデータに当てはめたグラフを図 2，図 3 に示す．そのときの残差平方和の対数値はそれぞれ 12.1957, 6.43325 となる．

表 3 前進差分を用いたパラメータ推定結果 DS-1.

	$\hat{A}$	$\hat{B}$	$\hat{c}$	$\hat{d}$
初期値	4.49023	0.0618383	0.0577542	0.147779
最適値	10.9903	0.112919	2.24715	-1.32898

表 4 前進差分を用いたパラメータ推定結果 DS-2.

	$\hat{A}$	$\hat{B}$	$\hat{c}$	$\hat{d}$
初期値	22.6558	0.270122	3.88851	-4.75046
最適値	6.02232	0.0698728	0.177778	-0.551277

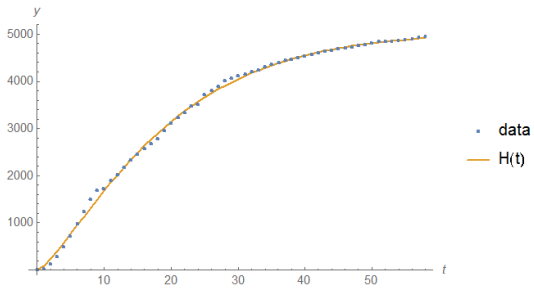

図 2　DS-1 の $H(t)$.

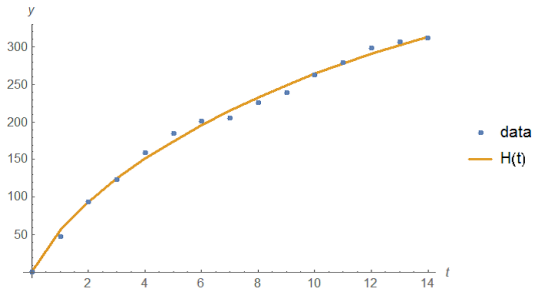

図 3　DS-2 の $H(t)$.

b)　後退差分の場合

$z(c,d,t_i)$ を後退差分を用いて，以下のように変形する．

$$z(c,d,t_i) = \log\left[\frac{y_i - y_{i-1}}{t_i - t_{i-1}} / t_i^c / y_i^d\right]. \qquad (15)$$

同様に式 (15) を用いて，式 (9)～(11) を計算し，パラメータを推定した結果を表 5, 表 6 に示す．

表 5 後退差分を用いたパラメータ推定結果 DS-1.

	$\hat{A}$	$\hat{B}$	$\hat{c}$	$\hat{d}$
初期値	-0.695989	0.0316029	-1.4843	1.29696
最適値	N/A	N/A	N/A	N/A

表 6 後退差分を用いたパラメータ推定結果 DS-2.

	$\hat{A}$	$\hat{B}$	$\hat{c}$	$\hat{d}$
初期値	-17.3404	0.16465	-5.17333	5.39424
最適値	N/A	N/A	N/A	N/A

後退差分では，1 回目の誤差最小化の段階で推定した c の値が -1 を下回ってしまったために，2 段階目でパラメータを推定できなくなったと考えられる．

c)　中心差分の場合

$z(c,d,t_i)$ を中心差分を用いて，以下のように変形する．ただし，$i = n$ の場合は中心差分を行えないので，代わりに後退差分を用いる．

$$z(c,d,t_i) = \begin{cases} \log\left[\frac{1}{2}\left(\frac{y_{i+1}-y_i}{t_{i+1}-t_i} + \frac{y_i-y_{i-1}}{t_i-t_{i-1}}\right)/t_i^c/y_i^d\right] & (1 \le i < n) \\ \log\left[\frac{y_n-y_{n-1}}{t_n-t_{n-1}}/t_n^c/y_n^d\right] & (i = n) \end{cases}. \qquad (16)$$

同様に式 (16) を用いて，式 (9)～(11) を計算し，パラメータを推定した結果を表 7, 表 8 に，推定したパラメータを $H(t)$ に代入してデータに当てはめたグラフを図 4，図 5 に示す．そのときの残差平方和の対数値は前進差分のときと同じくそれぞれ 12.1957, 6.43325 となる．

表 7 中心差分を用いたパラメータ推定結果 DS-1.

	$\hat{A}$	$\hat{B}$	$\hat{c}$	$\hat{d}$
初期値	2.65401	0.0492696	-0.532285	0.572545
最適値	10.9903	0.112919	2.24715	-1.32898

表 8 中心差分を用いたパラメータ推定結果 DS-2.

	$\hat{A}$	$\hat{B}$	$\hat{c}$	$\hat{d}$
初期値	5.22987	0.105936	0.0843702	-0.330734
最適値	6.02242	0.0698733	0.177796	-0.551301

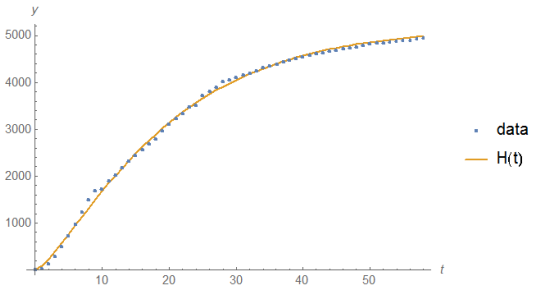

図 4　DS-1 の $H(t)$.

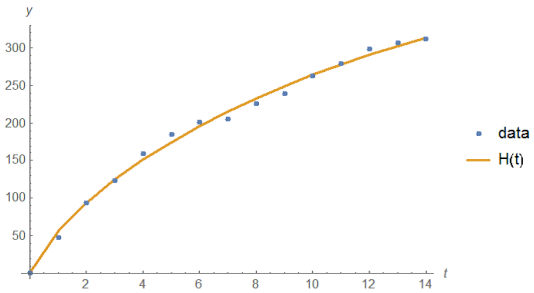

図 5　DS-2 の $H(t)$.

2 段階目でのパラメータの推定値は，DS-1，DS-2 ともに前進差分と中心差分に大きな差は見られなかった．データ数が大きい場合，前進差分の方が中心差分よりも 1 段階目でのパラメータの推定値と 2 段階目のパラメータの推定値との乖離が小さいので，以降は前進差分を用いて近似を行い，数値計算を行う．

(2)　$H(t)$ の 95% 信頼区間と予測区間の導出

DS-1 に対して，$H(t)$ の 95% 信頼区間と予測区間を導出する．そのためには，$\hat{A} - \hat{B}t$ の 95% 信頼区間と予測区間を導出する必要がある．

a)　$\hat{A} - \hat{B}t_i$ の 95% 信頼区間と予測区間

式 (12)，式 (13) を用いて，DS-1 の回帰の信頼区間，予測区間を求めていくが，回帰直線のパラメータを用いて信頼区間と予測区間を求め $H(t)$ に変換した場合，図 1 で示したように $H(t)$ がうまく当てはまっていないため，正しい $H(t)$ の

信頼区間と予測区間を求めることができない．よって，$H(t)$ がデータに対してうまく当てはまっている式 (11) で求めたパラメータを用いて $H(t)$ の信頼区間と予測区間を求める．その際，$\hat{A},\hat{B}$ が単回帰モデルでパラメータ推定する場合と同じ正規分布に従うと仮定する．また，残差 ε_i が正規分布に従っていなければ求めることができないため，まず残差が正規分布に従っているかどうかを調べる．

$\varepsilon_i = z(\hat{c},\hat{d},t_i) - (\hat{A} - \hat{B}t_i) \quad (i = 1,2,\cdots 58)$ を求め，ε_i に対して，帰無仮説 H_0 : “残差は正規分布に従っている”，対立仮説 H_1 : “残差は正規分布に従っていない” で Kolmogorov-Smirnov 検定 (以下 K-S 検定) を行った結果，p 値が 0 となり，帰無仮説が棄却されるため，信頼区間と予測区間を求めることができない．ここで，$z(\hat{c},\hat{d},t_i)$ と $\hat{A} - \hat{B}t_i$ のグラフを図 6，残差 ε_i が正規分布に従っていると仮定した場合の $\hat{A} - \hat{B}t_i$ の 95% 信頼区間と予測区間のグラフを図 7 に示す．

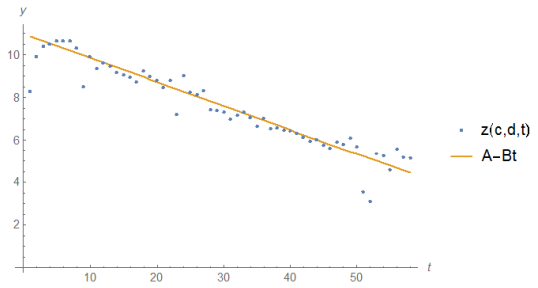

図 6 DS-1 の $z(\hat{c},\hat{d},t_i)$ と $\hat{A} - \hat{B}t_i$.

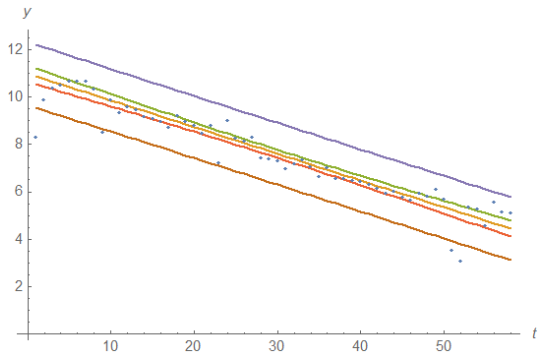

図 7 DS-1 の $\hat{A} - \hat{B}t_i$ の 95% 信頼区間と予測区間.

図 7 を見ると，予測区間の外に $z(\hat{c},\hat{d},t_i)$ が何点か存在していることがわかる．予測区間の外にある 4 点を外れ値として扱い，$z(\hat{c},\hat{d},t_i)$ の値を便宜上 $\hat{A} - \hat{B}t_i$ で置き換え，再度 $\varepsilon_i = z(\hat{c},\hat{d},t_i) - (\hat{A} - \hat{B}t_i) \quad (i = 1,2,\cdots n)$ を求め，ε_i に対して，帰無仮説 H_0 : “残差は正規分布に従っている”，対立仮説 H_1 : “残差は正規分布に従っていない” で K-S 検定を行った結果，p 値が 0.240658 となり，帰無仮説を棄却できないので残差 ε_i は正規分布に従っていないとは言えない．したがって，新しく求めた残差が正規分布に従っていると仮定して，95% 信頼区間と予測区間を求める．

DS-1 の $\bar{t}, S_t, V_e$ を求めると，$\bar{t} = 29.5, S_t = 16254.5, V_e = 0.135087$ となり，$t_{0.025}(56) = 2.00324$ なので，代入すると修正後の $\hat{A} - \hat{B}t_i$ の 95% 信頼区間は

$$\hat{A} + \hat{B}t_i - h_1(t_i) \le \hat{A} - \hat{B}t_i \le \hat{A} + \hat{B}t_i + h_1(t_i), \qquad (17)$$

と求まる．ここで，

$$h_1(t_i) = 2.00324\sqrt{\left(\frac{1}{58} + \frac{(t_i - 29.5)^2}{16254.5}\right) \times 0.135087},$$

である．同様に予測区間は

$$\hat{A} + \hat{B}t_i - h_2(t_i) \le \hat{A} - \hat{B}t_i \le \hat{A} + \hat{B}t_i + h_2(t_i), \qquad (18)$$

と求まる．ここで，

$$h_2(t_i) = 2.00324\sqrt{\left(1 + \frac{1}{58} + \frac{(t_i - 29.5)^2}{16254.5}\right) \times 0.135087},$$

である．式 (17)，式 (18) のグラフを図 8 に示す．

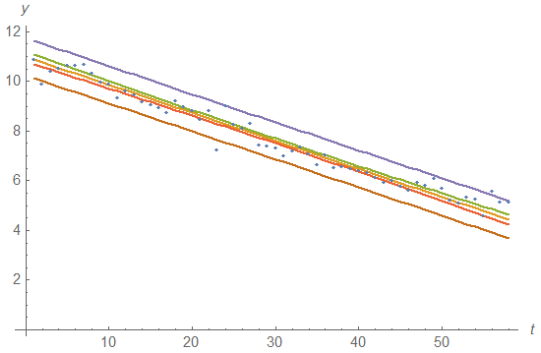

図 8 修正後の DS-1 の $\hat{A} - \hat{B}t_i$ の 95% 信頼区間と予測区間.

b) $H(t)$ **の信頼区間と予測区間**

式 (17)，式 (18) を用いて $H(t)$ の 95% 信頼区間と予測区間を導出する．式 (17)，式 (18) をそれぞれ式 (2) の右辺に代入した微分方程式の解がそれぞれ $H(t)$ の 95% 信頼区間と予測区間になるが，この微分方程式は解くことができないので，$\hat{A} - \hat{B}t_i$ と信頼区間，予測区間の差を $\hat{A}$ に足し，その値を $H(t)$ の A に代入して $H(t_i)$ を求め，求めた点を滑らかにつなぐことで信頼区間，予測区間を描く．$H(t_i)$ の 95% 信頼区間は

$$H(t_i, \hat{A} - h_1(t_i), \hat{B}, \hat{c}, \hat{d}) \le H(t_i) \le H(t_i, \hat{A} + h_1(t_i), \hat{B}, \hat{c}, \hat{d}),$$

となり，95% 信頼区間のグラフを図 9 に示す．同様に，$H(t_i)$

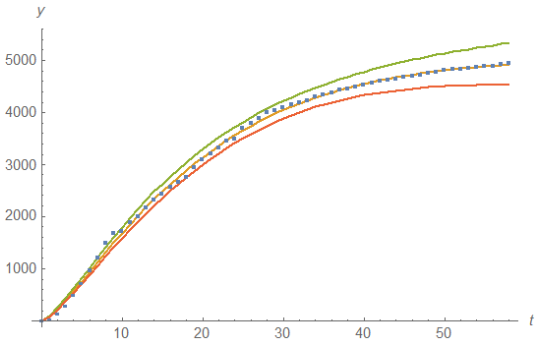

図 9 DS-1 の $H(t)$ の 95% 信頼区間.

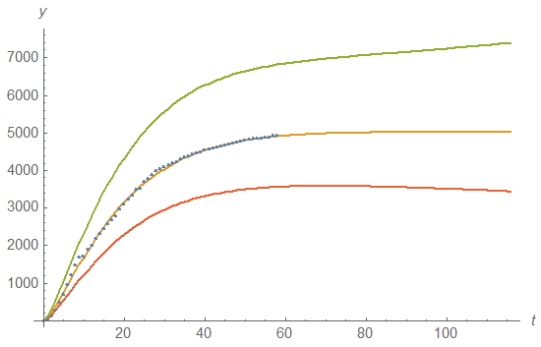

図 10 DS-1 の $H(t)$ の 95% 予測区間.

の 95% 予測区間は

$$H(t_i, \hat{A} - h_2(t_i), \hat{B}, \hat{c}, \hat{d}) \leq H(t_i) \leq H(t_i, \hat{A} + h_2(t_i), \hat{B}, \hat{c}, \hat{d}),$$

となり，95% 予測区間のグラフを図 10 に示す．本研究では，扱うデータが累積発見バグ数であることから既に観測されている値よりも下回ることがないため，上側の信頼区間と予測区間がより重要だと考えられる．よって，以降は上側の区間の曲線に着目していく．

(3) 累積バグ発見数の将来予測

バグが将来的にいくつ発見されるかを早い段階で正確に予測することが求められる．本研究では，$m < n$ として，DS-1 の (t_i, y_i) $(i = 1, \cdots, m)$ を用いて，$H(t)$ を推定し，(t_i, y_i) $(i = m+1, \cdots, m)$ の累積バグ数を予測できるかを検証する．$m = 28$ までデータの個数を減らして計算することができ，そのときの，$\hat{A}, \hat{B}, \hat{c}, \hat{d}$ の値を表 9，$H(t)$ のグラフを図 11 に示す．そのときの全体の誤差二乗和の対数値は 14.0706 で，$m + 1$ から n までの誤差二乗和の対数値は 13.9268 となる．

表 9 パラメータ推定結果 DS-1.($m = 28$)

	$\hat{A}$	$\hat{B}$	$\hat{c}$	$\hat{d}$
初期値	3.7406	0.0160338	-0.654148	0.399157
最適値	4.35605	0.0563806	0.172277	0.128651

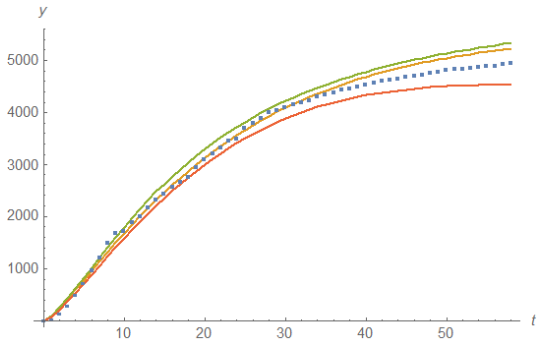

図 11 DS-1 の $H(t)$ の 95% 信頼区間 ($m = 28$).

図 11 を見ると，$H(t)$ は 95% 信頼区間の内側に収まっているが，予測した部分のデータ点のずれが大きくなっているため正確に推定できているとは言い難い．ただ，データ点の上側にずれているため，下側にずれているよりかは大きな問題ではないと考えられる．

8. 考察

先行研究 [4] と DS-1 で精度の比較を行うと，誤差二乗和の対数値は先行研究のモデルが 12.346 で本研究のモデルが 12.1957，AIC [7] は先行研究のモデルが 486.609 で本研究のモデルが 479.843 と先行研究のモデルよりも本研究で構築したモデルの方が良い結果となった．これは，パラメータの数を一つ増やしたことで，先行研究では扱えてなかった成長曲線を一般化することができ，よりデータに対してうまく当てはめることができたのが要因だと考えられる．また，早期推定に関しても先行研究が 35 日目までのデータを使って $H(t)$ を求めた際に 95% 信頼区間から外れたのに対し，本研究では 28 日目までのデータで $H(t)$ を求めても 95% 信頼区間から外れなかったため先行研究よりも精度のいいモデルを構築できたと考えられる．

9. おわりに

本研究では，Box-Cox 変換を用いて信頼性の分野でよく用いられる成長曲線を一般化した．また，一般化したモデルを用いて 95% 信頼区間を導出し，バグ数の将来予測を行った．データを用いた数値実験の結果として，先行研究よりも精度の良いモデルであることが示された．今後の課題としては，信頼区間と予測区間を求める際にパラメータが回帰直線の最小二乗法の結果で得られる分布に従うという仮定について，理論的に正しいことなのか検証することである．

参考文献

[1] M.Kimura, “A Study on Bootstrap Confidence Intervals of Software Reliability Measures Based on an Incomplete Gamma Function Model”, Advanced Reliability Modeling II Reliability Testing and Improvement, pp. 419-426 (2006).
[2] G.E.P. Box and D.R. Cox, “An analysis of transformations”, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 26, No. 2, pp. 211–252 (1964).
[3] 三浦武，ソフトウェアの品質評価法，日科技連出版社 (1981)
[4] 藤澤峻作，「成長曲線の一般化と差分方程式による解析に関する研究」，法政大学大学院理工学研究科修士論文 (2013).
[5] 山田茂, ソフトウェア信頼性評価技術, HBJ 出版局 (1989).
[6] 田口玄一，真壁肇，古林隆，森雅夫，確率・統計，日本規格協会 (1981).
[7] 鈴木義一郎，情報量規準による統計解析入門，講談社 (1995).